NIPT time point selection and abnormality determination of female fetuses via KMeans and random forest

Rongrong Zhao [1, #], Yifei Zhu [2, #], Jiayi Zhou [2, #], Dongfang Xie [3, *]

[1] Computing Science and Artificial Intelligence College, Suzhou City University, Suzhou, China, 215104

[2] School of Optical and Electronic Information, Suzhou City University, Suzhou, China, 215104

[3] Department of Basic Courses, Suzhou City University, Suzhou, China, 2151041

* Corresponding Author Email: xiedongfangkk2007@163.com

#These authors contributed equally.

Abstract. This paper explores non-invasive prenatal testing (NIPT), emphasizing the optimization of detection timing and the assessment of chromosomal abnormalities, utilizing data predominantly from pregnant women with high body mass index (BMI). The study begins by analyzing the relationship between Y-chromosome concentration in male fetuses and influencing factors, such as gestational age and BMI, through data preprocessing, visualization, Spearman correlation, and multiple regression modeling, which demonstrates significant linear associations. For male fetuses, it further applies K-means clustering based on BMI to group subjects, determining optimal detection timing via the 80th percentile of the first gestational week per group, with validity confirmed by ANOVA and silhouette scores. The research then extends to incorporate additional factors like height, age, and weight for refined grouping, while highlighting how measurement errors can distort rates and risk assessments, potentially complicating clinical decisions. Regarding female fetuses, the investigation develops a random forest model that integrates multiple features, including Z-scores of chromosomes 13, 18, 21, and X, GC content, read ratio, and BMI, to enhance the accuracy of aneuploidy detection beyond the limitations of single-feature approaches. Finally, the paper discusses the strengths and weaknesses of the proposed models, offering recommendations for their practical application and generalization.

Keywords: Multiple Regression Model, KMeans Clustering Model, ANOVA Test, Random Forest.

1. Introduction

Non-invasive prenatal testing (NIPT) is a prenatal testing technique that collects maternal blood, examines cell-free fetal DNA fragments, and analyzes for chromosomal abnormalities [1-3]. Its purpose is to determine the health of the fetus through early detection. Clinical experience suggests that there are three main types of chromosomal abnormalities in fetal malformations: Down syndrome, Edwards syndrome, and Patau syndrome. These abnormalities are determined by the presence of abnormal chromosome concentrations. The accuracy of NIPT is primarily determined by the fetal sex chromosome concentrations. Fetal sex chromosome concentrations are typically tested between 10 and 25 weeks of gestation. If the Y chromosome concentration in male fetuses is at or above 4% and the X chromosome concentration in female fetuses is normal, the NIPT result is considered generally accurate. Otherwise, accuracy cannot be guaranteed. Furthermore, early detection of unhealthy fetuses is crucial, as this risks shortening the treatment window. Early detection carries a lower risk, mid-term detection carries a higher risk, and late detection carries an extremely high risk [4-7].

Practice has shown that the Y chromosome concentration in male fetuses is closely correlated with gestational age and body mass index (BMI). Clinical practice typically groups pregnant women according to their BMI and determines the appropriate NIPT timing for each group. However, due to individual differences in age, BMI, and pregnancy status, applying simple empirical grouping and a uniform NIPT timing to all pregnant women can significantly impact test accuracy. Therefore, rationally grouping pregnant women based on BMI and determining the optimal NIPT timing for

each group can mitigate the potential risk of a shortened treatment window due to fetal health problems [8-10].

This modeling approach consists of the following steps:

(1). Developing a correlation model between fetal Y chromosome concentration and maternal indicators such as gestational age and BMI, and testing the model's significance.

(2). Stratifying BMI groups pregnant women with male fetuses based on the main factors influencing the time it takes for fetal Y chromosome concentration to reach the standard, determining the BMI range and optimal NIPT timing for each group to minimize potential risks, and analyzing the impact of testing error on the results.

(3). Stratifying BMI groups pregnant women with male fetuses, taking into account multiple factors, testing error, and the proportion of fetal Y chromosome concentrations reaching the standard, determining the optimal NIPT timing for each group to minimize potential risks, and analyzing the impact of testing error on the results.

4. Based on the aneuploidy of chromosomes 21, 18, and 13 in pregnant women with female fetuses listed in Appendix AB, a model for determining female fetal abnormalities was constructed by comprehensively considering factors such as the Z value, GC content, number of read segments and related ratios, and BMI of the X chromosome and the above chromosomes.

2. Analysis and model assumptions

2.1. Analysis

The data for this paper are from https://www.mcm.edu.cn/index_cn.html. First, we needed to clarify the relationship between fetal Y chromosome concentration and gestational age and BMI. We preprocessed the raw data. Next, we plotted Y chromosome concentration against gestational age and BMI to observe the distribution trends and preliminarily determine the overall pattern of Y chromosome concentration changes with gestational age and BMI. We then calculated the Spearman correlation coefficient between Y chromosome concentration, gestational age, and BMI.

Based on this, we constructed a multivariate regression model with Y chromosome concentration as the dependent variable and gestational age and BMI as independent variables. Our goal was to rationally stratify pregnant women carrying male fetuses based on BMI, determine the optimal NIPT timing for each group to minimize potential risk, and analyze the impact of testing errors. We first preprocessed the data to screen for valid samples and remove outliers.

Then, we used a clustering algorithm and the elbow rule to scientifically stratify BMI. Finally, we determined the optimal testing timing based on the success rate. Pregnant women carrying male fetuses were rationally stratified based on BMI, with each group assigned a BMI range and optimal NIPT timing to minimize potential risk. The impact of test error on the results was also analyzed. The effects of factors such as height, weight, age, gestational age, and BMI on Y chromosome concentration and their impact on error need to be comprehensively considered. This study employed a two-stage model: K-means clustering and a risk quantification function. The K-means clustering stage divided pregnant women into K groups, eliminating grouping issues caused by multifactorial differences. The risk quantification function stage constructed a comprehensive risk function to determine the optimal NIPT timing with the lowest combined risk score.

Finally, for multi-feature classification prediction, the core approach was to combine information on chromosomes 21, 18, 13, and X of female fetuses with BMI and GC content. A random forest model was constructed and applied to analyze data on chromosomal abnormalities in female fetuses. First, the raw data was filtered to identify valid samples with a GC content of 40% to 60%, a raw read count greater than 0, and a duplicate read ratio between 0 and 1. Chromosome aneuploidy results were then encoded. Subsequently, through comparative analysis of logistic regression, random forest and XGBoost models, the applicability of the random forest model was determined.

2.2. Model Assumptions

(1) Multiple test results for the same pregnant woman are independent of each other, and the errors follow a normal distribution.

(2) Unhealthy babies will not affect the time when the Y chromosome concentration reaches the standard.

(3) The height and weight of the pregnant woman remain unchanged during the short period of testing.

(4) The number of raw reads in the "Y chromosome concentration standard determination" only affects the test accuracy and does not change the time when the concentration reaches the standard.

3. Model building and solving

3.1. Multiple regression model

This study employed a multiple regression model. Its core principle was to fit the parametric relationship between the dependent variable (male fetal Y chromosome concentration) and the independent variables (gestational age, BMI, and its squared term) by minimizing the residual sum of squares (RSS). This model outputs the regression coefficients for each variable on Y concentration, and statistical significance of these relationships can be verified.

Step 1: Determine the form of the regression model based on the data analysis results.

The dependent variable Y concentration and the independent variables gestational age, BMI, and BMI squared term are all continuous, and marginal effects need to be quantified. The model was determined to be a multiple regression model with linear parameters and nonlinear transformations of the independent variables.

Step 2: Determine the coefficient that minimizes the residual sum of squares.

Let the residual of the i-th sample be $e_i = Y_i - \hat{Y}_i$ where Y_i is the observed Y concentration and $\hat{Y}_i$ is the predicted value. The objective function is to minimize the residual sum of squares:

$$\min_{\beta_0,\beta_1,\beta_2,\beta_3} RSS = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i2}^2)]^2 \quad (1)$$

Taking partial derivatives of RSS with respect to $\beta_0, \beta_1, \beta_2, \beta_3$ and setting them to 0, we obtain the equation system $(X^T X)\hat{\beta} = X^T Y$, and obtain the estimated value of the regression coefficient $\hat{\beta}$.

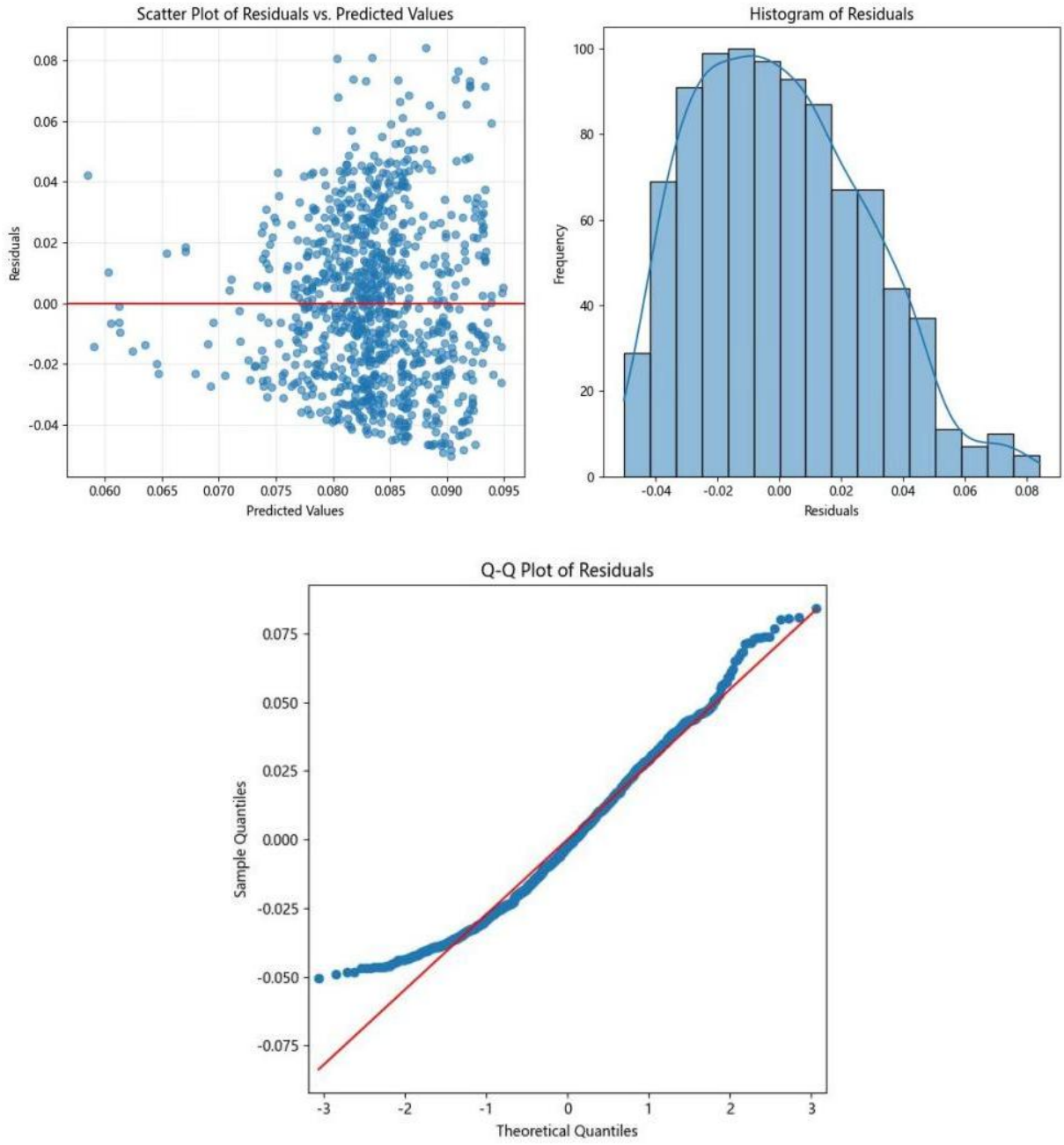

Figure 1. Multiple regression model solution results

This study established a multivariate regression model to investigate the relationship between Y chromosome concentration in male fetuses and gestational age and maternal BMI, using the ordinary least squares method. Figure 1 shows that, despite a low coefficient of determination ($R^2 = 0.038$), the overall linear relationship was significant ($F = 12.00$, $p < 0.001$). This indicates that, after controlling for other variables, the selected independent variables still have a statistically significant effect on Y chromosome concentration, albeit with a limited proportion of the variance explained. Gestational age showed a significant positive correlation with Y chromosome concentration. The effect of BMI on Y chromosome concentration followed a quadratic curve relationship, with an initial increase followed by a decrease.

3.2. KMeans algorithm

This paper uses the KMeans clustering model to group pregnant women's BMI data. The core of this algorithm is to achieve optimal sample partitioning by minimizing the within-cluster sum of squares (WCSS).

1. Model Construction:

The core of the KMeans clustering model is the within-cluster sum of squares (WCSS), which is mathematically expressed as:

$$inertia = \sum_{i=1}^{K} \sum_{x' \in C_i} ||x' - \mu_i'||^2 \quad (2)$$

To determine the optimal number of clusters K, the elbow rule is used:

(1) Experimental design: Set the number of candidate groups in the range K∈{2,2.5...5}, execute the KMeans clustering algorithm once for each K, and calculate the corresponding inertia value.

(2) Curve feature analysis: Draw the K-inertia curve. As K increases, the inertia value decreases monotonically; when K is small, the inertia value decreases significantly, indicating that increasing the number of groups can significantly improve the compactness of samples within the group; when K reaches a certain threshold, the rate of decrease of the inertia value slows down significantly, and the curve shows an "elbow" feature. The K corresponding to this elbow is the optimal value that takes into account both clustering quality and model simplicity.

(3) Result determination: Based on the comprehensive curve morphological characteristics, K=3 is the optimal number of clusters.

3. Calculation of the optimal NIPT time point and BMI interval:

(1) Determination of the optimal NIPT time point: The 80% percentile of the "first target gestational age" of each group is used as the theoretical optimal time point. This percentile can ensure that at least 80% of pregnant women have a Y chromosome concentration that reaches or exceeds the 4% accuracy standard when tested at this time point. It includes the principle of early detection and achieves an optimal balance between accuracy and risk control.

(2) Determination of the BMI interval: For each cluster obtained by clustering, the minimum and maximum BMI values of the samples in the cluster are calculated respectively, and the two are used as the interval endpoints to construct the BMI interval [min (BMI), max (BMI)] of each group, thereby obtaining the BMI distribution range of different groups.

Table 1. kmean calculation results

	BMI Range	Optimal NIPT Time
Group 1	[29.1, 33.3]	16
Group 2	[26.6, 31.9]	13.2
Group 3	[31.9, 38.5]	13.6

From the comparison between prediction data and actual data shown in Table 1, the BP neural network has better prediction performance and relatively small error, which can meet the demand completely, and has fast prediction speed and convenient operation.

3.3. Multi-factor clustering-risk optimization two-stage model

This study uses a two-stage model of multi-factor clustering and risk optimization. The core principle is to measure the similarity of samples within the group by minimizing the sum of squared errors within the cluster, obtain the initial grouping, and then quantify the risks at different gestational weeks to find the optimal detection time point.

Step 1 Verify whether the features are reasonable

Use the Pearson correlation coefficient to test the correlation between the features and the fetal Y chromosome concentration

Step 2 Use the KMeans clustering algorithm to group

(1) Use feature standardization to eliminate dimensional differences and obtain standardized values. For example, if the unit of height in cm is different from the unit of weight in kg, the formula is:

$$x'_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \tag{3}$$

(2) Use the method of minimizing the sum of squared errors within a cluster to measure the similarity of samples within a group. The formula is:

$$Inertia = \sum_{k=1}^{K} \sum_{x \in C_k} \| x' - \mu'_k \|^2 \tag{4}$$

Running the code shows that the optimal number of clusters, k=4, and the silhouette coefficient is 0.262. This indicates that the grouping effectively distinguishes pregnant women of different body types and ages. For example, Group 2 represents the high BMI group, with a BMI range of 31.7 to 41.1, while Group 3 represents the low BMI group, with a BMI range of 27.6 to 35.7. Optimal timing is determined as follows:

The low BMI group (Group 3) has the earliest optimal timing (11 weeks): Because this group's Y concentration reaches the standard slowly, the robust standard-reaching rate at the early stage (11 weeks) is only 0.5. The medium-to-high BMI group (Group 1) has the latest optimal timing (20 weeks): Because this group's standard-reaching rate increases significantly with gestational age, while clinical risk is higher in the late stage, the superior standard-reaching rate (0.859) partially offsets this risk, resulting in the lowest risk score at 20 weeks. The high BMI group (Group 0) has the earliest optimal timing (20 weeks): This balances the early standard-reaching rate of 0.859 with clinical risk, avoiding high risk in the late stage.

3.4. Random Forest Algorithm

This study employed a multi-model comparison, analyzing logistic regression, random forest, and XGBoost models. The results showed that the logistic regression model had an accuracy of 0.5689, while the random forest model achieved an accuracy of 0.9589. The XGBoost model also achieved an accuracy of around 0.9589. The logistic regression model was eliminated first. Although the random forest and XGBoost models achieved similar accuracy, the random forest model offered more intuitive interpretability, while the XGBoost model offered more complex interpretability. Given the need to investigate methods for identifying abnormalities in female fetuses, the random forest model was selected after comprehensive consideration.

The raw data contained a significantly higher proportion of "normal" samples, far exceeding other samples. However, there were only five samples with "Abnormality 21." Therefore, directly training the model would have resulted in a significant underdiagnosis rate. Therefore, we employed the SMOTE oversampling algorithm to expand the number of samples in each class to 341 (the same number as the "normal" samples), balancing the training set across classes and ensuring the model could fairly learn features from both abnormal and normal conditions.The visualization of a single decision tree (with a depth limit of 3) in the random forest model clearly illustrates the model's hierarchical decision logic for determining "female fetal chromosomal abnormalities." This prioritizes sequencing data quality by focusing on core abnormality metrics such as the "chromosome Z score," while also considering interference from maternal physiological characteristics. After model training, the visualization and quantitative metrics are shown in the figure below. The confusion matrix verifies that the random forest algorithm achieves a near-zero missed diagnosis rate for core screening categories such as "chromosome 13, 18, 21, and multiple abnormalities," providing valuable guidance for early detection of fetal anomalies using NIPT. The random forest algorithm's feature importance ranking not only focuses on core testing metrics such as "chromosome Z-score," but also considers factors influencing test accuracy, such as maternal BMI and sequencing quality. This demonstrates the strong interpretability and medical plausibility of the model's decision-making.

4. Conclusions

This paper effectively demonstrates the value of data-driven models, specifically KMeans clustering and Random Forest, for optimizing NIPT protocols. The KMeans model offers a superior, data-centric alternative to traditional clinical grouping for determining the optimal testing timepoint for male fetuses. By automatically creating clusters based on the principle of high intra-cluster similarity, it ensures a more precise BMI stratification, minimizing the risk of inaccurate timing due to high variance within groups. Its computational efficiency makes it practical for large datasets, and its flexibility allows for the incorporation of multi-dimensional features like BMI, age, and height to enhance personalization, as required. For the critical task of female fetal abnormality detection, the Random Forest model proved highly effective, integrating multiple genetic and maternal features to achieve high accuracy and provide interpretable decision logic. However, the KMeans model's static nature, which overlooks the temporal variable of gestational age, and its assumption of linear feature independence are notable limitations. Despite this, the timing optimization is particularly suitable for regions with high obesity rates, while the robust classification model can be adopted by tertiary hospitals to improve the early and accurate detection of chromosomal aneuploidies. Future research could focus on developing dynamic clustering models that incorporate temporal variables like gestational age to track stratification changes over time. Furthermore, exploring the integration of Random Forest with advanced deep learning architectures could enhance the detection of complex, non-linear relationships for a broader range of genetic abnormalities.

Acknowledgements

The authors gratefully acknowledge the financial support from The Natural Science Foundation of The Jiangsu Higher Education Institutions of China (Grant 23KJD110002) funds.

References

[1] Shi Weihui, Xu Chenming. Application value of non-invasive prenatal testing in the diagnosis of maternal complications and comorbidities in obstetrics [J]. Obstetrics and Gynecology Genetics Center, Obstetrics and Gynecology Hospital Affiliated to Fudan University, 2025, 41 (08): 617-619

[2] Jiang Liya, Lu Shaokan, Du Jiaen, et al. Development and application of non-invasive prenatal testing technology [J]. Clinical Medical Research and Practice, 2025, 10 (23): 191-194.

[3] PENG H, WANG D, GUO F, et al. Prenatal diagnosis of imprinted associated chromosome abnormalities identified by noninvasive prenatal testing (NIPT)[J]. Scientific Reports, 2025, 15 (1): 12830–12830.

[4] POULTON A, HUI L. Noninvasive prenatal testing: an overview [J]. Australian Prescriber, 2025, 48 (2): 47–53.

[5] GAZDARICA J, FORGACOVA N, SLADECEK T, et al. Insights into non-informative results from non-invasive prenatal screening through gestational age, maternal BMI, and age analyses [J]. PLOS ONE, 2024, 19 (3): e0280858.

[6] JUUL L A, HARTWIG T S, AMBYE L, et al. Noninvasive prenatal testing and maternal obesity: A review [J]. Acta Obstetricia et Gynecologica Scandinavica, 2020, 99 (6): 744–750.

[7] ABEDALTHAGAFI M, BAWAZEER S, FAWAZ R I, et al. Non-invasive prenatal testing: a revolutionary journey in prenatal testing [J]. Frontiers in Medicine, 2023, 10: 1265090.

[8] TAMAK S, ESLAMI Y, DA CUNHA C. Validation of multidimensional performance assessment models using hierarchical clustering [J]. Expert Systems with Applications, 2025, 290: 128446.

[9] SADVAKASSOVA L, KURMANGALI Z, BELOUSOV V, et al. The effectiveness of non-invasive prenatal test technology and the prenatal screening algorithm based on various methods for determining foetal aneuploidy [J]. Journal of the Turkish German Gynecological Association, 2023, 24 (3): 152–158.

[10] KATRACHOURAS A, KONTOS H, KONIS K, et al. Early Non-Invasive Prenatal Testing at 6–9 Weeks of Gestation [J]. Genes, 2024, 15 (7): 895.